STA130 – Winter 2021

Week 11 Problem Set — Part 1 Sample Answers

N. Moon & S. Caetano

Instructions

How do I hand in these problems for the 11:59 a.m. ET, April 1st deadline?

Your complete .Rmd file that you create for this problem set AND the resulting .pdf (i.e., the one you 'Knit to PDF' from your .Rmd file) must be uploaded into a Quercus assignment (link:https://q.utoronto.ca/courses/206597/assignments/578479) by 11:59 a.m. ET, on April 1st. Late problem sets or problems submitted another way (e.g., by email) are *not* accepted.

Problem set grading

There are two parts to your problem set. One is largely R-based with short written answers and the other is more focused on writing. We recommend you use a word processing software like Microsoft Word to check for grammar errors in your written work. Note: there can be issues copying from Word to R Markdown so it may be easier to write in this file first and then copy the text to Word. Then you can make any changes flagged in Word directly in this file.

Part 1

Question 1

(Note: Question 1 also provides some revision of topics learned in previous weeks, as you prepare to submit your project and the final assessment)

Lumosity is a brain training app thought to help cognitive skills - for example, memory, reasoning and focus. A large randomized trial was conducted to evaluate the impact of Lumosity training on cognitive skills. The study and results are presented in: Hardy, JL, Nelson, RA, Thomason, ME, Sternberg, DA, Katovich, K, Farzin, F, et al. (2015) "Enhancing Cognitive Abilities with Comprehensive Training: A Large, Online, Randomized, Active-Controlled Trial". *PLoS ONE* 10(9): e0134467. doi:10.1371/journal.pone.0134467.

Thousands of participants were recruited from Lumosity's free users (i.e., people who set up free Lumosity accounts but did not pay for full access) and randomly assigned to either:

- Lumosity training (Treatment) - complete Lumosity training online for approximately 15 minutes at a time, at least 5 times a week for 10 weeks, or
- Crossword puzzles (Control) - complete crossword puzzles online for approximately 15 minutes at a time, at least 5 times a week for 10 weeks.

The main measure of cognitive skills was called the Grand Index (GI) Score; higher values mean better cognitive skills. The cognitive skills of the participants who completed the study were scored before and after the 10-week study period. We will store data on the improvement (i.e., after-before) in GI Scores (GI_improve) for the 5045 Lumosity users who participated in the study as well as several other the variables that may be useful are in a data frame called `study_dat`.

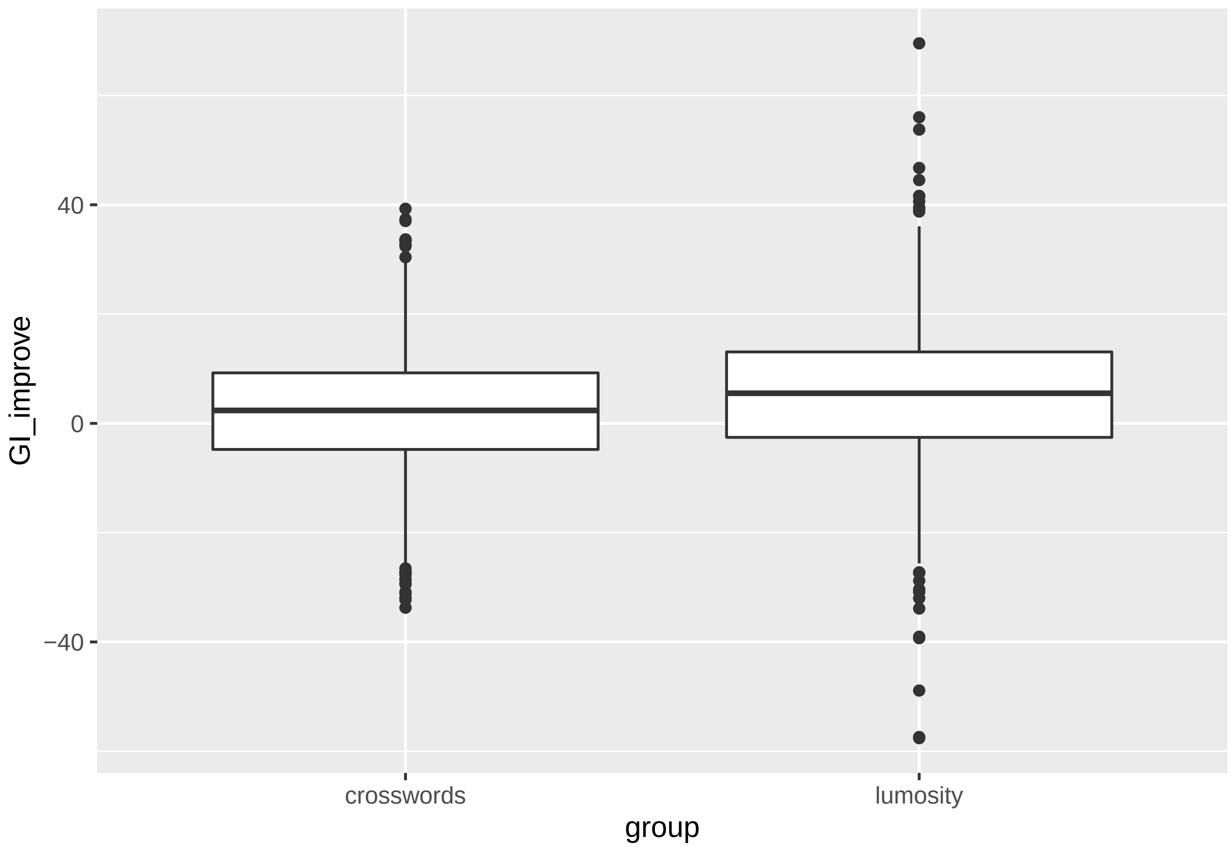
```r
library(tidyverse)

study_dat<-read_csv("lumosity_study_data.csv")
glimpse(study_dat)
```

```
## Rows: 5,045
## Columns: 6
## $ participant_id     <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 1...
## $ group              <chr> "crosswords", "lumosity", "crosswords", "crosswo...
## $ age_round          <dbl> 51, 24, 19, 21, 31, 25, 45, 27, 40, 50, 46, 52, ...
## $ GI_improve         <dbl> -5.992334, 9.030539, -15.030917, -9.265834, 3.18...
## $ concentration_post <dbl> 1, 3, 5, 2, 4, 1, 4, 5, 5, 4, 4, 4, 5, 4, 4, 2, ...
## $ active_days        <dbl> 56, 54, 6, 21, 43, 15, 0, 40, 39, 69, 64, 32, 21...
```

Let's consider how Grand Index score improvements vary by type of online training:

```r
ggplot(study_dat, aes(x=group, y=GI_improve)) + geom_boxplot()
```



```
group_by(study_dat, group) %>%
  summarise(mean = mean(GI_improve),
          sd = sd(GI_improve),
          n = n())
```

```
## # A tibble: 2 x 4
##   group       mean    sd     n
## * <chr>      <dbl> <dbl> <int>
## 1 crosswords  2.14  10.6  2378
## 2 lumosity    5.24  12.0  2667
```

A hypothesis test on two groups can be conducted to compare mean Grand Index score improvements after online training with Lumosity and crosswords. (This might take a few moments to run.)

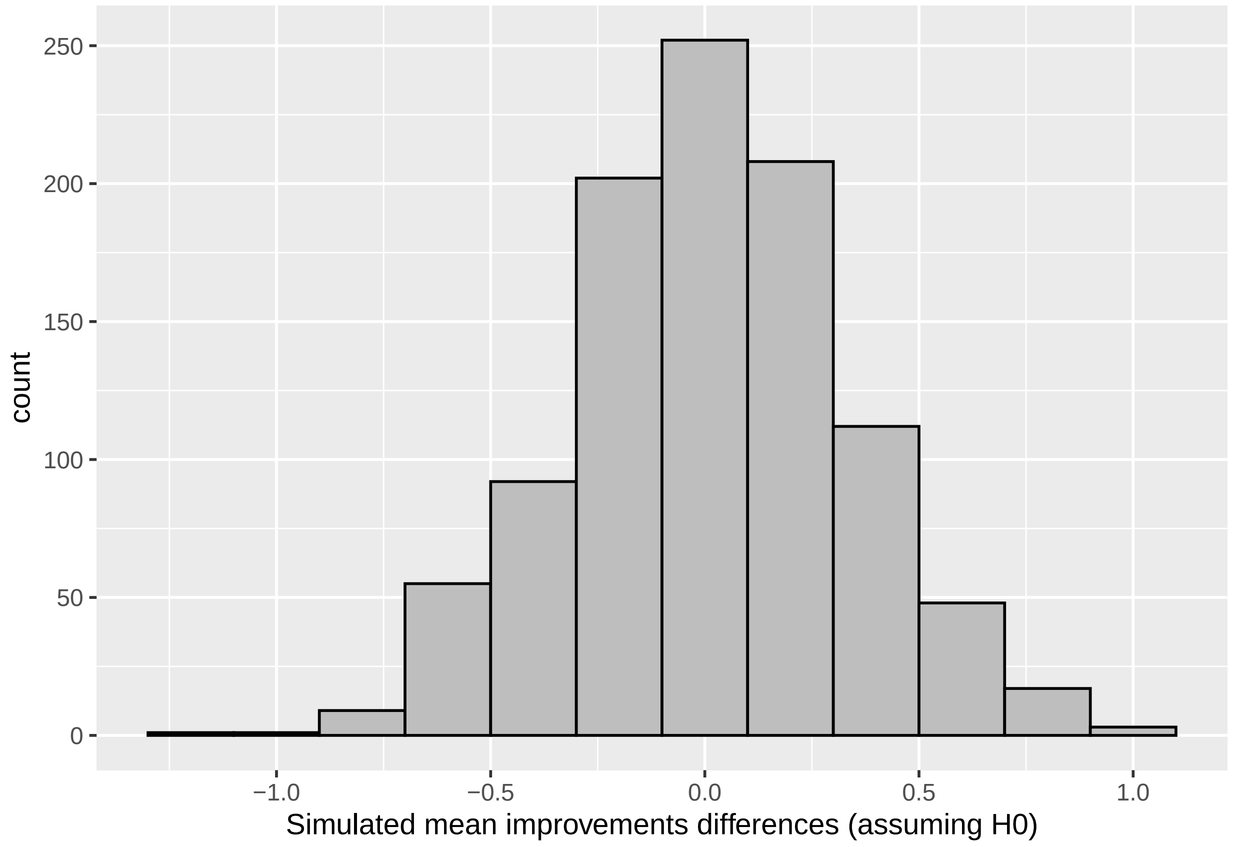
```
# compute test statistic
test_stat<-as.numeric(study_dat %>%
  group_by(group) %>%
  summarise(means = mean(GI_improve), .groups='drop') %>%
    #.groups='drop' is included to avoid a warning message being
    # printed, but doesn't change behaviour
  summarise(value = diff(means)))

# conduct randomization test
simulated_values <- rep(NA, 1000)
```

```
for (i in 1:1000) {
  sim <- study_dat %>% mutate(group = sample(group))
  sim_value <- sim %>%
    group_by(group) %>%
    summarise(means = mean(GI_improve), .groups=’drop’) %>%
    summarise(value = diff(means))
  simulated_values[i] <- as.numeric(sim_value)
}

sim <- tibble(mean_diff = simulated_values)

ggplot(sim, aes(x=mean_diff)) +
  geom_histogram(col="black",fill="gray", binwidth=0.2) +
  labs(x = "Simulated mean improvements differences (assuming H0)")
```



```
sim <- tibble(mean_diff = simulated_values)
sim %>%
  filter(mean_diff >= abs(test_stat) |
           mean_diff <= -1*abs(test_stat)) %>%
  summarise(p_value = n() / 1000)
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1       0
```

The estimated p-value based on this randomization test is 0 so there is very strong evidence against the hypothesis that the mean Grand Index Score improvement is the same for those training using the Lumosity app and those completing online crossword puzzles.

(a) Consider using a simple linear regression model instead to test for a difference in mean Grand Index score improvements for those who train using Lumosity and those who complete online crossword puzzles.

(i) Write down the appropriate regression model. Be sure to define any terms you use.

(ii) State the hypotheses to compare mean `GI_improve` when training using Lumosity and online crossword puzzles based on the model you specified in the previous part of this question.

(iii) Use R to fit this model and interpret the estimated regression coefficients.

(iv) We'll assume for the purposes of this question that the necessary assumptions for valid inference based on the tests conducted on the regression parameters by R are reasonable here.

Interpret the p-value of this test to compare the mean improvement for Lumosity versus crossword puzzles. How does it compare to the p-value estimated using the randomization test earlier in this question? Is this surprising? Why or why not.

(b) Consider the study design used by Hardy et al. (2015).

(i) What type of study did Hardy et al. (2015) conduct? Use vocabulary from the course and justify your answer based on how the researchers did the study.

(ii) Can we conclude that Lumosity training leads to more improvement in cognitive skills than completing crossword puzzles online based on these results? Explain your answer.

(iii) As reported in Hardy et al. (2015), 9919 participants consented to participate and were randomly assigned to a training type. However, only 5045 study participants actually completed the study. The dataset only included data on study participants who completed the study. How might this limit our conclusions?

(c) Perhaps `age` of the user is related to cognitive improvement as well.

(i) Do you think user ages would be different between the Lumosity group and crossword groups? Why or why not?

(ii) Produce an appropriate data summary to see if ages of the users differ for the Lumosity and crossword groups. Interpret your summary and comment on how this compares to your prediction about how ages would compare in *c(i)*.

```r
#Write your code to produce the summary in here.
```

(iv) Suppose that there was a big difference in the age distributions of the two treatment groups - for example, suppose that younger users were much more likely to drop out of the Lumosity group than to drop out of the crossword puzzle group, and so the mean age of individuals who completed the study was 50 for the Lumosity group and 38 for the crossword puzzle group. How might this limit (if at all) the conclusions of the analysis?

Note: The following questions are 'short answer', you only need to write 1 to 3 sentences.

Question 2

(Modern Data Science with R Exercise, 2nd edition, Section 8.12 Problem 8) A data analyst received permission to post a data set that was scraped from a social media site. The full data set included name, screen name, email address, geographic location, IP (internet protocol) address, demographic profiles, and preferences for relationships. Why might it be problematic to post a deidentified form of this data set where name and email address were removed?

Question 3

(Modern Data Science with R Exercise, 2nd edition, Section 8.12 Problem 5) A reporter carried out a clinical trial of chocolate where a small number of overweight subjects who had received medical clearance were randomized to either eat dark chocolate or not to eat dark chocolate. They were followed for a period and their change in weight was recorded from baseline until the end of the study. More than a dozen outcomes were recorded and one proved to be significantly different in the treatment group than the outcome. This study was publicized and received coverage from a number of magazines and television programs. Outline the ethical considerations that arise in this situation.

Part 2

Watch the following short video on the Tuskeege syphilis trial (link: [https://www.youtube.com/watch?v=afwK2CVpc9E&feature=share&fbclid=IwAR3r8TMnlNGIObYtp7NWg-krrBFMe6Ui9k4zqbcsGflh8g2Jxv_ymlMFkk8]).

If you are not able to access the video using the first link, a version has been uploaded to mymedia **here** (link:[https://mymedia.library.utoronto.ca/play/a80e488a0d9bd09193c9833336515ca5]).

If possible, please watch the original video on Youtube to support the creators of this content. When watching the video, consider what are the main ethical concerns of the Tuskeege syphillis trial.

Your assignment is as follows: Identify at least 2 ethical concerns and describe them. Then, explain how you could conduct a similar trial today while avoiding some of the same ethical pitfalls that you identified from the original study.

As this is the last writing prompt of the semester, you will be REQUIRED to submit a 4 to 5 minute video OR voice clip. Do not feel the need to do tons of 'takes'. Rather, you can repeat yourself if you make a mistake, or feel you are unclear. This is not meant to be an additional burden, but rather to provide you with the opportunity to practice your oral communication skills and get a break from writing.

You might be wondering how can I record this? One way to do this would be to schedule a Zoom meeting and record yourself in it. You can record the video to the cloud, or even directly on your computer! There will be many file types, including a video version, and one that is just a voice recording.

You MUST upload a link (aka a URL) for your TA to watch your video. You can do this by uploading your video to mymedia. Alternatively, you can provide the zoom link from your recording. ONLY links will be reviewed by the TA. If you upload a video or voice file it will not be accepted.

If you are looking for more ideas of how to record yourself for this assignment or run into issues on how to upload your assignment, please post to Piazza.